International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-9, Issue-6; Jun, 2022
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.96.2

The Quality of Drinkable Water using Machine Learning Techniques

Osim Kumar Pal

Department of Electrical & Electronics Engineering, American International University-Bangladesh, Bangladesh
Email: osimkpal@gmail.com

Received: 04 May 2022,

Received in revised form: 24 May 2022,

Accepted: 31 May 2022,

Available online: 06 Jun 2022

***Keywords*— Artificial intelligence, Artificial Neural Network, Big data, Prediction model, Water quality.**

***Abstract*—** *Predicting potable water quality is more effective for water management and water pollution prevention. Polluted water causes serious waterborne illnesses and poses a threat to human health. Predicting the quality of drinkable water may reduce the incidence of water-related diseases. The latest machine learning approach has shown promising predictive accuracy for water quality. This research uses five different learning algorithms to determine drinking water quality. First, data is gathered from public sources and presented in accordance with World Health Organization (WHO) water quality standards. Several parameters, including hardness, conductivity, pH, organic carbon, solids, and others, are essential for predicting water quality. Second, Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), Deep Neural Network (DNN), and Gaussian Nave Bayes are used to estimate the quality of the drinking water. The conventional laboratory technique for assessing water quality is time-consuming and sometimes costly. The algorithms proposed in this work can predict drinking water quality within a short period of time. ANN has 99 percent height accuracy with a training error of 0.75 percent during the training period. RF has an F1 score of 87.86% and a prediction accuracy of 82.45%. An Artificial Neural Network (ANN) predicted height with an F1 score of 96.51 percent in this study. Using an extended data set could improve how well predictions are made and help stop waterborne diseases in the long run.*

I. INTRODUCTION

1.1 Context

The Drinkable water quality prediction is essential to ensure safe public health. It is a very much serious issue for a person to survive healthy life. Polluted drinking water can cause various kinds of health diseases. According to the survey, almost 3,575,000 people are died every year due to water-related diseases [1]. Predicting drinkable water is difficult for those countries that have limited drinkable water sources. In the industrial revolution, chemical dust causes the most water pollution.

There is various kind of predicting methods to predict the drinkable water. Among those, neural network [2], gray theory [3], statistical analysis, and chaos theory [2] are the most useable techniques. For ideal model designing, statistical analysis is very much superior. For better prediction and research, a neural network delivers better performance [2]. Drinking water quality mainly depends on essential measures, such as pH, hardness, sulfate, organic carbon, turbidity, and a few more [4]. Machine learning techniques show significant prediction results in water quality prediction. Artificial neural network (ANN), Convolutional neural network (CNN),

Deep neural network (DNN), Random Forest (RF), Support vector machine (SVM) are the most popular machine learning algorithm for prediction [5].

1.2 Problem

Water pollution is becoming the most severe human concern affecting water quality. Various human activities render water unsafe for drinking and domestic usage. The primary causes of water pollution are chemical fertilizers and pesticides that enter rivers and streams as untreated wastewater and industrial effluents that run near cities and lowlands. Polluted water increases certain waterborne infectious illnesses, causing some severe diseases.

The issues that this study intends to solve are outlined below:

a) misconception of WHO guidelines on drinkable water parameters;

b) the lengthy clinical process of drinkable water prediction;

c) lack of uses of machine learning on water quality prediction;

d) key awareness factor that are unknown to rural people.

1.3 Objectives

The primary goal of this project is to develop a computationally competent and robust approach for estimating drinkable water quality characteristics to reduce the effort and expense associated with measuring those parameters. The WHO standards on drinkable water and the awareness factors that may reduce water pollution will be reviewed. This study is about underground water in the Bogura District of northern Bangladesh, where the quality of the water is always changing.

II. REVIEW OF RELATED WORKS

A hybrid decision tree-based machine learning model was proposed to predict the water quality with 1875 data. In the evaluation process, six water quality parameters were used to predict the water quality. Extreme gradient boosting (XGBoost) and RF algorithms were applied that includes complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) along with six different algorithms. At first, raw statical data was collected. After CEEMDAN distribution, XGBoost and RF algorithms were applied in data distribution section. When training was completed, it shows the water quality along with prediction error [6].

A machine learning model was proposed with RF, Decision Tree (DT) and Deep Cascade Forest (DCF). The first step of the prediction model was data processing. Data samples were divided into suitable and unsuitable section at data processing unit. After that, system calculated the water quality parameters for irrigation. Water quality was predicted by six levels of measure. Data was collected from Bouregreg watershed (9000 km2) located in the middle of Morocco. Data was divided into 75 percent for training and 25 percent for testing. In the data normalization and model building unit, system predict the water quality by data splitting.[5].

An author presented a data intelligence model for water quality index prediction. Support vector regression (SVR), adaptive neuro-fuzzy inference system (ANFIS), Back propagation neural network (BPNN) and one multilinear regression (MLR) algorithms are applied for prediction. The author collected the data from Jumna, the major tributary of the Ganga River. The length of the river is 1400 km. [7].

A hybrid machine learning approach was suggested for water quality prediction. RF, reduced error pruning tree (REPT), and twelve different algorithms were applied to analyze the water quality. The author divided the methodology into two sections are data collection and preparation. Eleven water quality indicators were applied to identify the water quality. In the model evaluation, the author took coefficient of determination (R2), mean absolute error (MAE), root-mean-square deviation, the percentage of bias (PBIAS), percent of relative error index (PREI), and Nash-Sutcliffe efficiency (NSE) for the performance measure of different algorithms. [8].

III. PROPOSED METHODOLOGY

3.1 Introduction

Machine learning algorithms, classification algorithms, and regression algorithms all improve daily in our contemporary age, producing improved results. The most often used classification algorithms are ANN, CNN, DNN, DT and RF [5]. Using factors such as pH, conductivity, hardness, and so on, this proposed model predicts whether or not the water is safe to drink.

Numerous methods using activation functions are utilized in data processing and learning. RF, SVM, ANN, DNN and Gaussian Naïve Bayes are the suggested prediction algorithms in this proposed work.

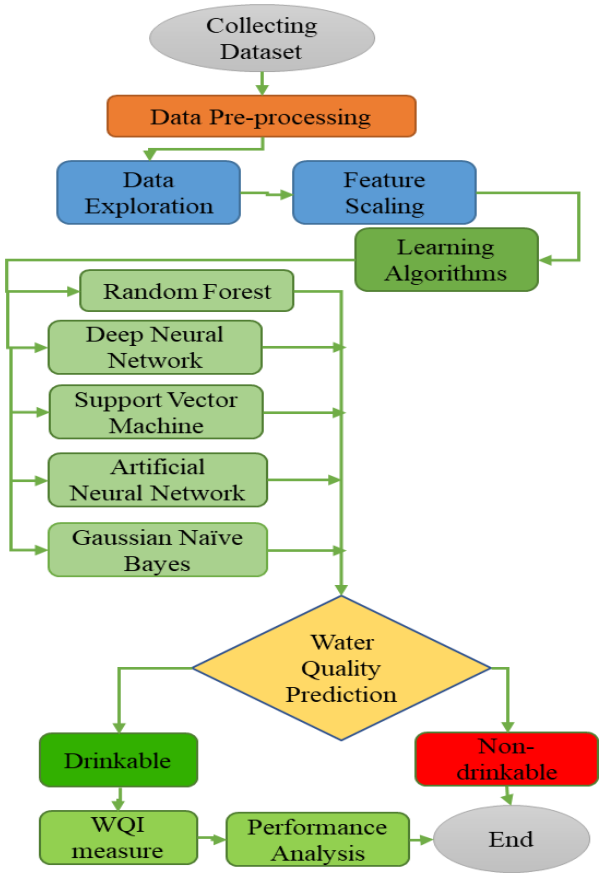

Fig. 1: Framework of proposed model

To begin, data are collected and data are distributed according to ten measurements shows in Fig. 1. Then, algorithms are developed according to literature analysis. After that, five distinct classifiers will be built to categorize the data and predict the class. Finally, the suggested study presents prediction findings together with a performance analysis. Performance analysis identifies the optimal method.

3.2 Dataset

This research is used a dataset from Department of Public Health Engineering (Rajshahi Branch, Bangladesh). It constituted 3276 samples. The dataset includes the following key metrics: pH, hardness, solids (total dissolved solids - TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and portability. The standard data rate established by the International Water Association ensures the quality of drinking water in Bangladesh [9].

3.3 Data Processing

The computation step is critical in data processing for improving data quality. In this step, data exploration and feature scaling being determined using the dataset's most important parameters. The samples of water were then categorized based on the WQI values.

3.4 Water quality classification and index calculation

WQI measures water quality by factoring in factors that affect WQ [10].

$$WQI = \frac{\sum_{i=1}^{N} q_i * w_i}{\sum_{i=1}^{N} w_i} \quad (1)$$

The WQI was determined using the formula:

$$q_i = 100 * \left(\frac{V_i - V_{ideal}}{S_i - V_{ideal}}\right) \quad (2)$$

Here,

N = No. of parameters

q_i = quality rating scale

w_i = weight of each parameter

K = proportionality constant

The proposed model is evaluated in this study using a public dataset and ten critical water quality indicators.

Table. 1: Drinkable Water Quality Standards

Parameters	Unit	Standards
pH		6.5-8.5
Hardness	mg/L	300
Solids (TDS)	ppm	<20000
Chloramines	mg/L	<4
Sulfate	mg/L	<250
Conductivity	µS/cm	<400
Organic Carbon	ppm	<25
Trihalomethanes	µg/liter	<37
Turbidity	NTU	<5

Note: World Health Organization water quality standard

In Table. 1. It shows the standard value of water quality index and those measurements are provided by World Health Organization (WHO) [11].

3.5 Machine learning algorithms

3.5.1 Random Forest

Random forest is a Classification Algorithm extensively utilized in Multiclass applications. It constructs classification trees from several samples and uses their majority vote for classification and average for regression. Many of the most significant characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous variables (as in regression) and categorical variables (as in classification). It outperforms

other algorithms in categorization tasks. Random forest actually uses to methods: Bagging and Boosting [12].

Some important feature that makes RF more accurate.

1. Diversity
2. Immune to the curse of dimensionality
3. Parallelization
4. Train-Test split
5. Stability [13]

In regression problems, the mean squared error (MSE) rate is an important parameter in the RF. For calculating the value of MSE [14],

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \quad (3)$$

Here,

N= No. of the total data points.

f_i= Return value from the proposed model.

y_i= Data point's actual value.

3.5.2 Deep neural network

A deep neural network is much more complex than the first. It can understand voice instructions, identify sound and images, conduct an expert assessment, and a variety of other tasks that involve foresight, creativity, and analytics. Only the human brain is capable of such things. Unlike feed-forward networks (FFNs), deep neural networks (DNNs) include connections between layers that are only one-way and can only send data forward. The results are produced via deep classification with knowledge datasets, with "what we want" defined through the hidden layer. An FFNN's taste is like a memory trace [15].

DNN has 4 layers of operation

1. Dataset
2. Local Receptive Fields
3. Sharing Weights
4. Pooling Layer

The deep neural network addresses the issue on a larger scale and may make judgements or make predictions based on the data provided and the intended outcome. Without a large quantity of labeled data, a deep neural network can solve a problem [15].

3.5.3 Support vector machine

The term "Support Vector Machine" (SVM) refers to a supervised machine learning method that may be used to solve classification and regression problems. It is, however, mostly employed to solve categorization issues. The SVM method displays each data item as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the coordinate value [16]. To compute the norm of a vector, use the Euclidean norm formula [17].

$$x = (x_1, x_2, \ldots, x_n) \quad (4)$$

If it defines x = (x_1, x_2) and w = (a, −1)

$$w * x + b = 0 \quad (5)$$

The hyperplane may then be used to create predictions. The hypothesis function h is defined as follows [17]:

$$h(x) = \begin{cases} +1 \; if \; m * x + b \geq 0 \\ -1 \; if \; m * x + b < 0 \end{cases} \quad (6)$$

3.5.4 Gaussian naive bayes

The Naive Bayes method is a probabilistic machine learning technique that may be used to a broad range of classification problems. Filtering spam, categorizing documents, and predicting sentiment are examples of common uses. The term naive refers to the assumption that the characteristics that make up the model are unrelated to one another. That is, altering the value of one feature has no direct impact on the value of the other characteristics utilized in the algorithm [18]. The term naive refers to the assumption that the characteristics that make up the model are unrelated to one another. To calculate the mean and variance of X, the formula is [19],

$$P(X|Y = c) = \frac{1}{\sqrt{2\pi\sigma^2 c}} * e^{-(x-\mu_c)^2 / 2\sigma^2 c} \quad (7)$$

Replacing the appropriate probability density of a normal distribution and name it the Gaussian Naive Bayes if it assumes the X's follow a Normal (aka Gaussian) Distribution, which is quite frequent [19].

3.5.5 Artificial neural network

The phrase "artificial neural network" refers to a sub-field of artificial intelligence influenced by biology and patterned after the brain. A computer network based on biological neural networks that build the structure of the human brain is known as an artificial neural network. Artificial neural networks, like human brains, contain neurons that are coupled to each other at different levels of the networks. Nodes are the name for these neurons [20].

3.6 Data distribution analysis

This study project includes ten measurements. Throughout the data distribution process, each statistic is shown individually to provide context for the drinkable water standard. pH is a unit of measurement that is used to indicate the acidity or basicity of an aqueous solution. In water, it indicates the alkaline measure. The WHO

recommends a pH range of 6.5 to 8.5 as the highest acceptable level [11]. Water hardness is the quantity of dissolved calcium and magnesium present in water that is measured as "water hardness." Hard water has a high

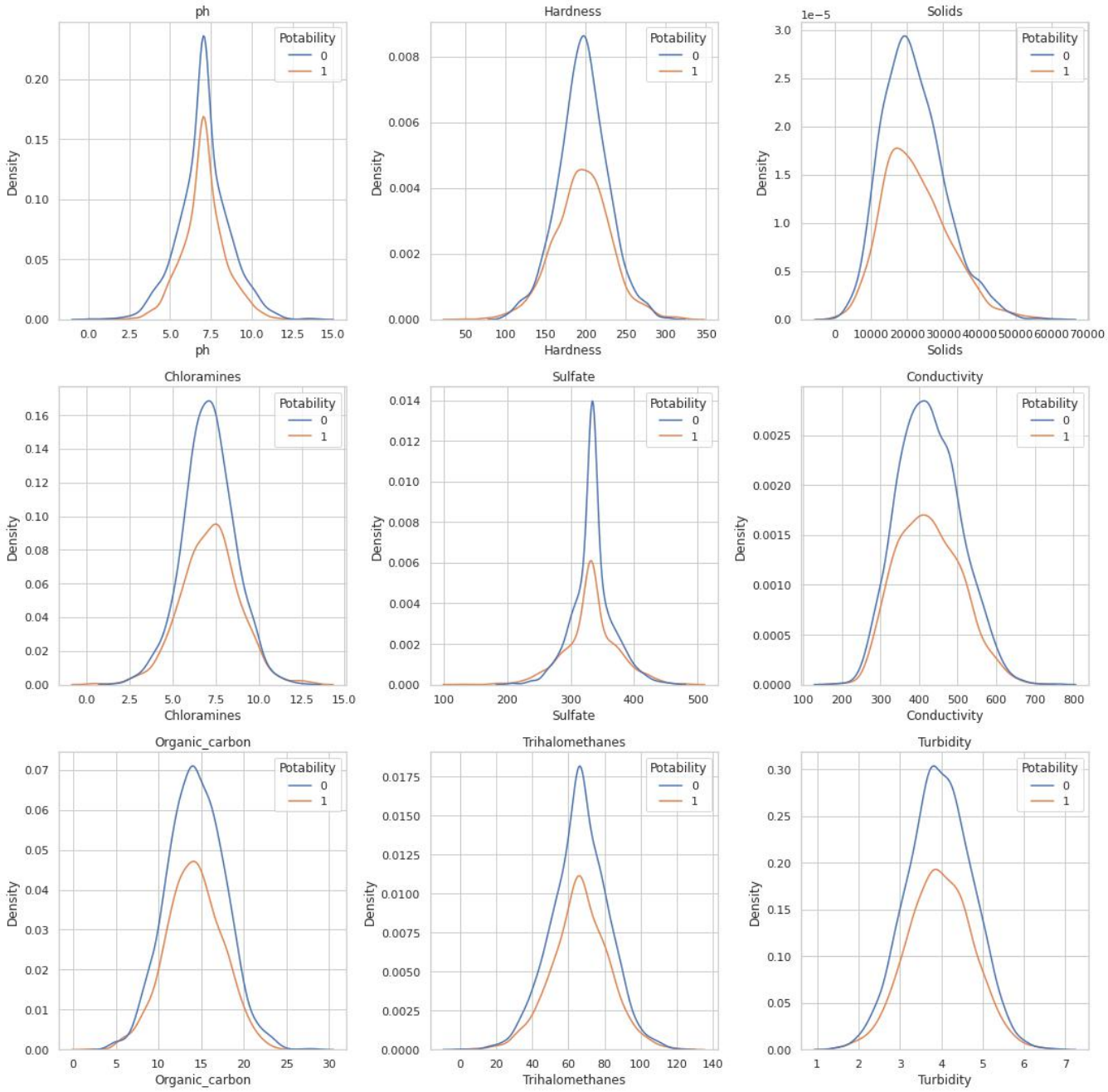

Fig. 2: Distribution of measuring parameters in terms of potability

concentration of dissolved minerals, primarily calcium and magnesium, and should be avoided. Standard water hardness by WHO is 300 mg/L [11].

Solids (Total Disable Solids) refers to the inorganic salts and trace quantities of organic materials that are present in solution in water. Calcium, magnesium, sodium, and potassium cations are frequently present, as well as carbonate, hydrogen carbonate, chloride, sulfate, and nitrate anions [21].

Chloramines is one of the important disinfectants that used in water potability. Fig. 2. shows the distribution of chloramines in the dataset. Under 4 mg/L is the standard rate of Chloramines in drinkable water. Sulfate may provide a bitter or medicinal flavor to water and has

laxative properties. Fig. 2. shows the distribution of sulfate in samples. The allowable sulfate in drinking water is under 250 mg/L [11].

In drinking water, electrical conductivity is a measurement of a solution's ionic mechanism that allows it to transfer electricity. Fig. 2. shows the distribution of conductivity. Based on WHO guidelines, the electrical conductivity value should not exceed 400 S/cm [11].

Organic carbon indicates organic matter in drinking water. It may have thousands of components, such as microscopic particles, dissolved macromolecules, colloids, and compounds [22]. The allowable rate of organic carbon in drinking water is lower than 25 ppm [11].

Trihalomethanes are disinfection byproducts formed when chlorine molecules combine with naturally existing substances in water. Trihalomethanes in drinking water have a standard value of 37 µg/liter [11]. They are colorless and will float on the surface of the water. The turbidity of water is determined by the amount of solid stuff suspended in it. The WHO recommends 5.00 NTU [11].

A correlation heatmap is a graphical representation of a correlation matrix that illustrates the relationships between different variables. The correlation coefficient may be anything between -1 and 1 [23]. Fig. 3 is a correlation heatmap created to illustrate the linear relationship between various variables on drinkable water quality in the dataset.

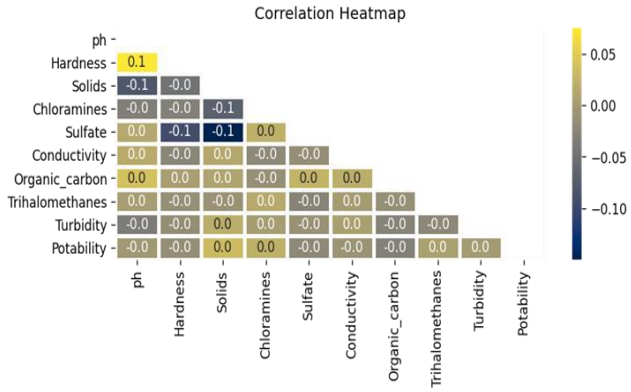

Fig. 3: Correlation Heatmap of ten variables in the experimental dataset

The dataset contains ten measurement parameters. When the value is 0.1 to 1, the correlation between two variables is considered to be positive. A positive value implies that when one variable rises, the other increases as well. Hardness and pH have a positive relationship shown in fig 10. A negative correlation between two variables is defined as a value of -1 to -0.1. A negative value implies that as one variable goes up, the other goes down. In this dataset, solids and sulfate have a relationship valued at -0.1. There is no connection between the two variables if the value is 0, which implies that the variables vary randomly [23]. Sulfate and pH do not correlate between them in this experimental dataset.

3.7 Performance parameters

The confusion matrix is one of the characteristics that properly represents the true performance of a classification model and may be used to monitor the system's performance. For assessment, the confusion matrix contains True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values.

The equation for calculating average accuracy,

$$Accuracy = \frac{TP+TN}{Total} \quad (8)$$

The equation for calculating Positive Predictive Value (PPV)/ Pression,

$$Pression = \frac{TP}{TP+FP} \quad (9)$$

For calculation True Positive Rate (TPR)/ Recall,

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

To calculate the F1 Score,

$$F1\ Score = \frac{PPV*TPR}{PPV+TPR} \quad (11)$$

F1 score is the performance measure over testing accuracy. It actually indicates that how stable the model is to predict the classes. If the F1 score is higher than the testing accuracy, then the system is more stable and accurate according to recall.

3.8 Model building

Random forest, deep neural network, support vector machine, gaussian naive bias and artificial neural network are applied for predicting the quality of drinking water. After the data visualization,

Table. 2: Model Building Parameter

Algorithms	Training accuracy (%)	Training error (%)
Random Forest	96	3.22
Deep Neural Network	94	4.5
Gaussian Naïve Bayes	98.59	1.77
Artificial Neural Network	99	0.75
Support Vector Machine	97	2.72

During model construction, the artificial neural network outperforms all other learning methods in terms of accuracy. Deep neural networks achieve 94 percent training accuracy with a 4.5 percent error rate, as shown in Table. 2. The artificial neural network has the lowest training error of any learning method, at 0.75 percent.

IV. SIMULATION RESULTS AND DISCUSSION

In this research work, PyCharm is used to analysis the data and predict the drinking water potability considering the ten measurements. During the prediction, WQI measured how safe the water was to drink. WQI informs that whether the water can be drunk or not based on WHO standards.

Table. 3: Drinkable and undrinkable of this water quality dataset in terms of WQI

Drinkable	Undrinkable
89.71 (%)	10.29 (%)

Figure. 3. shows that only a tiny amount of water is not safe to drink. When fertilizers, industrial waste, animal waste, chemical fertilizers, pesticides, and waste from landfills and septic systems leak into an aquifer, they pollute the groundwater [24]. Surface water gets polluted a lot when cities proliferate without planning. Good waste management, fair use of aquatic resources, and more public knowledge might reduce the pollution of drinking water. Five different machine learning algorithms are applied for prediction. Each of the algorithm is run for five times to get accurate and authentic performance measurement.

Table. 4: Performance analysis of proposed learning algorithms

Algorithms	Precision (%)	Recall	F1 Score	Testing Accuracy
Random Forest	91.65	84.38	87.86	82.45
Deep Neural Network	94	86.88	90.3	84.57
Gaussian Naïve Bayes	96.23	90.19	93.11	92.44
Artificial Neural Network	98.86	94.27	96.51	98.12
Support Vector Machine	92.55	90.72	91.63	93.17

Table. 4. shows that for the prediction of drinkable water, an artificial neural network obtains the highest accuracy of 98.12 percent with 96.51 percent F1 Score. According to the F1 score, the random forest, deep neural network, and Gaussian naive bias have higher prediction stability when compared to actual accuracy. Random forest shows the lower accuracy of 82.45 percent along with 87.86 percent F1 score. Overall artificial neural network shows highest accuracy among those five learning algorithms. Other algorithm shows better stability in prediction that observed by F1 Score measure.

V. CONCLUSION

Predicting drinkable water is essential for environmental preservation and pollution prevention. It is necessary to provide clean drinking water in order to maintain excellent public health. Drinking water from safe sources can ensure the potability of the water. It becomes difficult to predict drinkable water accurately. The ideal learning algorithm is needed to prevent prediction errors. An intelligent model based on five different machine learning algorithms may be used to predict the potability of drinking water based on 10 standard parameters such as pH, hardness, organic carbon, and other factors. In this current work, artificial neural network achieved 98.12 percent accuracy with 0.75 percent training error. In future, the proposed model will be implemented to predict and analysis of different region drinking water along with IoT based quality detection model.

REFERENCES

[1] Deaths from Dirty Water. Retrieved February 4, 2022, from https://www.theworldcounts.com/challenges/planet-earth/freshwater/deaths-from-dirty-water/story

[2] Wu, H., & Zhao, X. (2016). Prediction Simulation Study of Road Traffic Carbon Emission Based on Chaos Theory and Neural Network. International Journal of Smart Home, 10(7), 249-258. https://doi.org/10.14257/ijsh.2016.10.7.25 2

[3] Jin-qiang, C. (2019). Fault Prediction of a Transformer Bushing Based on Entropy Weight TOPSIS and Gray Theory. Computing in Science & Engineering, 21(6), 55-62. https://doi.org/10.1109/mcse.2018.2882357

[4] Zhang, Q., Xu, P., & Qian, H. (2020). Groundwater Quality Assessment Using Improved Water Quality Index (WQI) and Human Health Risk (HHR) Evaluation in a Semi-arid Region of Northwest China. Exposure and Health, 12(3), 487-500. https://doi.org/10.1007/s12403-020-00345-w

[5] Chen, K., Chen, H., & Zhou, C. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research, 171, 115454. https://doi.org/10.1016/j.watres.2019.115454

[6] Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality

prediction. Chemosphere, 249, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169

[7] Abba, S. I., & Pham, Q. B. (2020). Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. Environmental Science and Pollution Research, 27(33), 41524-41539. https://doi.org/10.1007/s11356-020-09689-x

[8] Bui, D. T., Khosravi, K., & Tiefenbacher, J. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

[9] Dphe.rajshahi.gov.bd. 2022. Department of Public Health Engineering, Rajshahi.. [online] Available at: <http://dphe.rajshahi.gov.bd/> [Accessed 3 June 2022].

[10] Akter, T., Jhohura, F. T., & Akter, F. (2016). Water Quality Index for measuring drinking water quality in rural Bangladesh: A cross-sectional study. Journal of Health, Population and Nutrition, 35(1). https://doi.org/10.1186/s41043-016-0041-5

[11] Guidelines for Drinking-water Quality, Fourth Edition. Retrieved April 3, 2022, from https://apps.who.int/iris/bitstream/handle/10665/44584/9789241548151_eng.pdf

[12] Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x 2

[13] Explaining Feature Importance by example of a Random Forest | by . Retrieved April 3, 2022, from https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

[14] Darmawan, M. F., Zainal Abidin, A. F., & Kasim, S. (2020). Random forest age estimation model based on length of left hand bone for Asian population. International Journal of Electrical and Computer Engineering (IJECE), 10(1), 549. https://doi.org/10.11591/ijece.v10i1.pp549-558

[15] Deep Neural Networks - KDnuggets. Retrieved April 3, 2022, from https://www.kdnuggets.com/2020/02/deep-neural-networks.html

[16] Understanding Support Vector Machine (SVM) algorithm from. Retrieved April 5, 2022, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[17] Mathematics Behind SVM | Math Behind Support Vector Machine. Retrieved April 6, 2022, from https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/

[18] Asriadie, M. S., Mubarok, M. S., & Adiwijaya,. (2018). Classifying emotion in Twitter using Bayesian network. Journal of Physics: Conference Series, 971, 012041. https://doi.org/10.1088/1742-6596/971/1/012041

[19] How Naive Bayes Algorithm Works? (with example and full code. Retrieved May 3, 2022, from https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/

[20] Artificial Neural Network Tutorial - Javatpoint. Retrieved April 12, 2022, from https://www.javatpoint.com/artificial-neural-network

[21] Solids Content of Wastewater and Manure | Oklahoma State University. Retrieved May 1, 2022, from https://extension.okstate.edu/fact-sheets/solids-content-of-wastewater-and-manure.html

[22] What is soil organic carbon? | Agriculture and Food. Retrieved April 13, 2022, from https://www.agric.wa.gov.au/measuring-and-assessing-soils/what-soil-organic-carbon

[23] Correlation Concepts, Matrix & Heatmap using Seaborn - Data. Retrieved May 1, 2022, from https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/

[24] Sources and Causes of Water Pollution That Affect Our Environment. Retrieved May 7, 2022, from https://www.conserve-energy-future.com/sources-and-causes-of-water-pollution.php